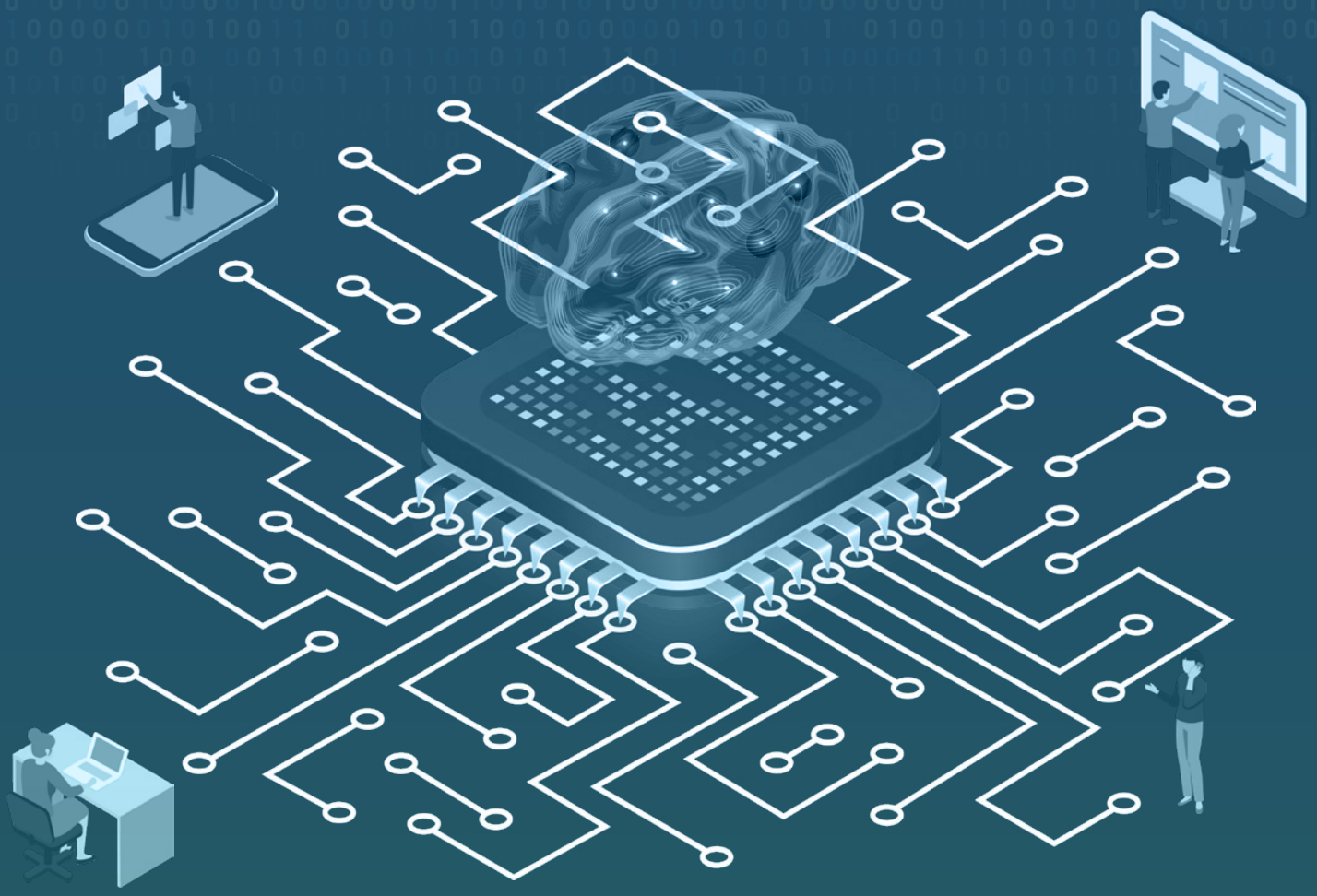


Algorithmic Audit for Decision-Making or Decision Support Systems

Matías Aránguiz Villagrán

IDB

Algorithmic Audit for Decision-Making or Decision Support Systems

Matías Aránguiz Villagrán
Professor and Deputy Director of Law, Science, and Technology Program
Pontificia Universidad Católica de Chile

March 2022

Acknowledgements
The author would like to thank Cristina Pombo for her comments and Sebastián Dueñas for the research during the writing of this document. Special thanks also to AGESIC, Uruguay's Agency for Electronic Government and the Information and Knowledge Society, especially to Maximiliano Maneiro for his review and comments, and to Eticas Consulting for the initial discussions on this document.

https://www.iadb.org/

Contents

1. Introduction

Decision-making is one of the core abilities of human beings. Deciding between more than one alternative allows us to discern and opt for better ways of doing things. Decision-making is a process through which a person weighs the available information and incorporates their previous experience to choose the option that, at the time, seems more convenient.

Daniel Kahneman, the winner of the 2002 Nobel Prize in economics, distinguishes two ways of thinking in human beings that operate in the decision-making: a first, fast, intuitive, and emotional system, and a second, slower, deliberative, and logical system. The first way is not always efficient, and the second, while delayed, allows reaching conclusions that incorporate more elements of analysis, a deeper level of reflection, and efficiency in decisions.

Kahneman shows that the way these two systems make decisions is complementary: speed is essential on some occasions, while complex and thorough analysis is critical on others.

Governments, corporations, institutions, and a broad range of groups make decisions that affect the lives of others (promotions, social benefits, criminal convictions, etc.). Therefore, decision-making should be a careful and comprehensive process that incorporates all the correct, updated, and relevant information, ensuring efficiency.

Given the number of affected individuals by governmental decisions and the degree of impact in their lives, these processes shall be conducted with special care and incorporating criteria of democratic participation and accountability.

1.1 What are automated decision support systems?

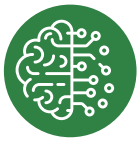
Automated decision support systems (ADS) are machine-based systems that can make predictions, recommendations, or decisions influencing real or virtual environments for a given set of human-defined objectives. These systems are designed to operate with varying levels of autonomy.[1]

1 OECD, Recommendation of the Council on Artificial Intelligence. Available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Over the last few years, ADS have grown exponentially in number and application areas. Currently, we interact with an increasing number of ADS, often without even noticing it. However, the lack of awareness about their use does not reduce the social risks if these systems are poorly designed or created without taking the necessary precautions.

If ADS are used with vulnerable groups or communities such as children, people with disabilities, and historically disadvantaged populations or at risk of exclusion,[2] having even greater foresight at the time of implementation will be necessary.

In this guide, ADS are categorized into two groups according to their degree of autonomy:



ADS in which the information generated by **automated learning models** is used as an input for **decision-making by an individual**

ADS in which final decisions and their resulting actions are made without **direct human intervention.**[3]

This guide is not intended exclusively as a practical instrument for identifying and mitigating risks or hazards that may not be apparent at first sight. It also serves as an instrument that helps raise awareness of the implications and consequences of implementing automated systems in making or supporting decisions that affect people's lives.

1.2 What is an audit for?

Usually, any system may fail or have risks undetected at first sight or whose relevance is overlooked due to the frequency with which certain processes are carried out. The more complex the system, the more likely errors will occur. Simultaneously, the system complexity often allows for greater adaptability to the reality on which they make predictions.

According to the ISO 19011 standard on "Guidelines for auditing management systems," an audit should be a systematic, independent, and documented process for obtaining evidence and evaluating it to determine the extent to which the criteria are fulfilled.[4]

An audit must incorporate the entity objectives, the protection of the interests and needs of its beneficiaries, collaborators, and other possible stakeholders, and the information safety and privacy requirements.[5] Accordingly, there are audits of varying types: accounting, legal, process, and IT audits, among others. The usefulness of these audits lies in the fact that they allow us to make an objective evaluation of the possible risks, quantify these, and prioritize their mitigation.

While audits have become a fundamental component in the growing field of algorithmic governance,[6] they are not sufficient to mitigate the impact of a system's implementation and execution; they are basically a process to determine compliance with some standards. However, audits play a crucial role in impact assessment and information gathering and availability, for the entity itself and regulatory entities, and also for those potentially affected and the society as a whole.

2 IDB. Ethical Assessment of AI for Actors within the Entrepreneurial Ecosystem: Application Guide. Available at: https://publications.iadb.org/publications/english/document/Ethical-Assessment-of-AI-for-Actors-within-the-Entrepreneurial-Ecosystem-Application-Guide.pdf

3 IDB: Responsible AI: Technical Manual: Life Cycle of Artificial Intelligence (IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial). Available at: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

4 ISO 19011-2018. Available at: https://www.iso.org/standard/70017.html

5 ISO 19011-2018. Available at: https://www.iso.org/standard/70017.html

6 Ada Lovelace Institute. Examining the Black Box: Tools for assessing algorithmic systems. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

1.3 What is an algorithmic audit?

An algorithmic audit is a study that seeks to evaluate ADS and its development process, including the design and data used to train the system. It also evaluates the impacts in terms of accuracy, algorithmic fairness,[7] bias, discrimination, privacy, and security, among others.[8]

Algorithmic audits can be conducted as measurement against certain standards (performance audits) or as a compliance analysis of particular standards (compliance audits) to produce recommendations on specific metrics.[9]

1.4 Who is this guide for?

This guide is addressed at policymakers in Latin America and the Caribbean and/or those responsible for leading ADS projects charged with mitigating the impacts produced by its use. This document is indented to serve as a guide to supervise these developments from pre-design stages to implementation, and possible adjustments and updates required for the appropriate use of the AI model. It is about supporting readers by guiding them on the need to audit artificial intelligence (AI) systems and indicating the elements to be considered while conducting the audit.

1.5 How to use this guide?

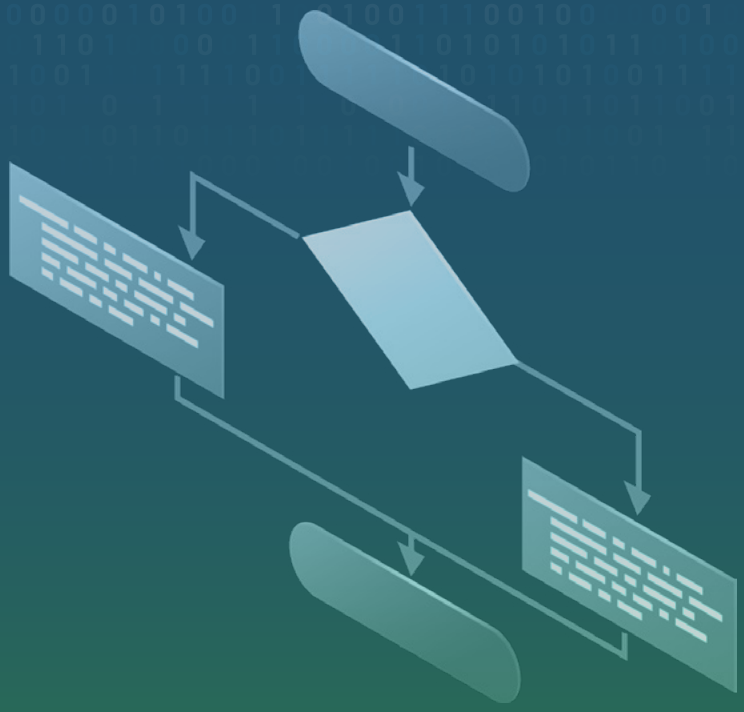
The purpose of this document is to introduce the reader to the subject using structured questions to decide on an audit implementation and the resulting process. The guide must be used as a support during the system life cycle:[10] from its conceptualization and design, to its use, and the corresponding accountability. It also includes references for those interested in gaining in-depth knowledge in specific topics, emphasizing those particularly relevant factors according to the type of entity, the source of the data, and/or the model used, among others.

7 In this context, algorithmic fairness is the feature of an algorithm, which, upon application, does not cause harm or discriminate against an individual or a group.

8 Algorithmic Accountability Act of 2019

9 INTOSAI. Performance Audit Principles. Available at: https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

10 Responsible use of AI for public policy: A project formulation guide. Available at: https://publications.iadb.org/publications/english/document/Responsible-use-of-AI-for-public-policy-Data-science-toolkit.pdf



2. The algorithmic audit

2.1 Why conduct an algorithmic audit?

With the widespread adoption of ADS in the public and private sectors, more and more dimensions in people's lives are under its influence. From the waiting time of public transportation users to the correct allocation of public services, in all of this, the goal is to achieve optimum benefit.

The implementation of automated systems leads to challenges, which are often not addressed in sufficient depth. Violations of fundamental rights for using personal data, unwanted discrimination by entities, and decisions that are difficult or even impossible to justify are just a few examples.

In light of the above, internal and external control and review measures are crucial, particularly in the public sector, where algorithmic audits are very useful. These are practical and effective processes conducted by third parties to guarantee that decisions are correct, observing ethical and technical considerations while respecting the rights of citizens.

While this is a subject whose regulation is still under discussion and development in multiple countries, the varying national policies related to AI and the numerous international guidelines reveal the need for adequate control mechanisms.

In public organizations, the performance of these audits allows to verify that the following purposes are fulfilled[11]:

11 IA Now Institute. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. Available at: https://ainowinstitute.org/aiareport2018.pdf

Respect the public's right to know which systems impact their lives by publicly listing and describing ADS that significantly affect individuals and communities.

Increase the internal capability of public agencies to evaluate the systems they build or procure and enable them to gain more experience on these tasks. Thus, they can anticipate issues that may arise from undesirable situations, such as incorrect benefits allocation or due process violations.

Ensure greater accountability in the use of ADS by designing a useful and ongoing method for third parties to review and assess these systems in a way that they can identify and solve or mitigate problems.

Ensure that the public has a meaningful opportunity to respond to and, if necessary, dispute the use of a given system or the guidelines used for its development by a public agency.

If, on the other hand, this type of audit is not conducted, the incorrect use of ADS might lead to a non-optimal use of resources or to triggering of cases of fundamental rights violations of various sectors of the public. Potential risks and damages are wide-ranging and often difficult to anticipate. There are basically two types: risks of inclusion (allocation of resources or benefits to those who do not need them) and risks of exclusion (deprivation of resources or benefits to people in need).

Uncontrolled implementation of ADS and/or lack of audits can also cause reputational damage to those implementing the system, rupturing thus the trust that society places in their diligent and correct actions. It can also lead to a generalized distrust of technology, making the public increasingly reluctant to use ADS in the public sector.

It is worth noting that other risks are inherent to AI tools development. These include, for example, overemphasis on specific performance metrics optimization to the detriment of the transparency and equity dimensions. Another clear risk from the lack of resources to develop models internally in the organizations that require them, which often decide to buy tools, albeit designed by third parties for multiple uses, are ultimately adapted for the purchaser's particular purpose. In addition, there is the risk that the system data may not be equally representative for all cases, creating a system

where inequality is the right thing. It can create difficulties in the model's adequacy, given its operational requirements and the regulatory demands for its operation.

Even so, algorithmic audits allow entities to satisfy the efficiency and effectiveness requirements that both public and private entities must comply with, either due to the existing regulations or citizen demands for transparency and efficiency.

2.2 Requirements to perform an algorithmic audit

The government entities/services or third parties can develop ADS internally. In the case of third parties, a product or service acquisition contract, a bid, or a direct purchase are alternatives to formalizing the development. When deciding on the acquisition of ADS, the institution must have someone responsible for the project, who can administer the contract and with some degree of technical knowledge to keep the development and implementation process under control.

When acquiring the service from a third party, the bidding or direct purchase contract shall include clauses that allow auditing the system. It is advisable not to limit the number of audits or their conditions in order not to prevent its performance when required. At the same time, the bid must clarify that audit can be done directly or by a third party on behalf of the public agency.

Some companies may be reluctant to be audited, especially if there is the possibility of information becoming public. Therefore, access to the source code of ADS should be required so that these can be audited ex-post. It is worth noting that supplier regulatory compliance rules can never be a constraint on conducting audits.

Also, the bid must specify that for the ADS service provision, submission of the technical documentation related to their development is required, including user manuals, policies, and technical descriptions of the training, design, and implementation processes. It is essential to keep a record allowing auditors to review the ADS.

2.3 When should an algorithmic audit be performed?

Algorithmic audits, unlike other evaluations, are conducted after the system has been implemented and is operating. Accordingly, ADS' design and development can be contrasted with the effects of its implementation, especially when there have already been cases of risks or damages.

Determining the exact moment to perform an algorithmic audit is not a minor exercise, usually, because the effects of ADS on the population are evident for the project lead only after damages have been produced. If they have not occurred, audits are strongly recommended at the end of the pilot period, with a controlled implementation on a sample of the total universe. For example, if ADS are intended to help qualify the socioeconomic risk of families to allocate social benefits, the pilot project should begin in a small city rather than in a region or the entire country.

At the end of the pilot project, i.e., after testing its implementation in a defined period and on a specific sample, it is advised to evaluate it. Scenarios can be considered before implementing the pilot, for example, in a simulation where possible errors can be detected based on different results scenarios. If the preliminary internal evaluation by the ADS development team reveals the existence of complications, the audit should be then performed.

Periodic performance of these audits is highly advisable, especially in constantly changing social contexts and/or systems whose operation may evolve according to a greater quantity and variety of data used, among others. The periodicity of these audits should also consider the system's risk of error. In case of errors or adverse effects, a periodic audit becomes imperative.

Systems criticality

Criticality refers to the importance and risk of ADS in their design and implementation. Importance means the function that ADS have within a process chain that feeds other systems with the information produced or their role in a defined situation. Some critical systems directly assign rights, aids, or subsidies, and others operate in naturally sensitive areas, such as national defense, health, or the prison system.

Risk means the possibility that during their use, ADS may produce errors that could cause harm to the population involved. Such would be the case of ADS that decide on the release of a defendant, the allocation of resources for social purposes, or the response to conflict situations. At least three elements should be considered to analyze the damage: (i) the probability of damage occurring is usually measured by determining how accurate the system is in accomplishing its task, or establishing how often it is wrong; (ii) the depth of the impact, meaning the slight or significant consequences the error may have (the most serious are those errors that inflict damage to life, freedom, or the property of an individual or a social group, or those depriving them of a public service or aid essential to their survival), and (iii) the distribution of the error, this is when the error made by ADS affects more than one sub-group of the population than others, as has happened with low-income groups, immigrants, and racial minorities. A frequent example is facial recognition systems that perform optimally with lighter-skinned people and usually make more mistakes with darker-skinned people. Discrimination against such subgroups causes unfair allocations or classifications and produces significant unease in the affected individuals.

When analyzing the information on sets of elements, namely importance and risk, it is possible to determine the criticality of a system. The more critical it is, the greater the precautions are required. Systems may generate such a risk that their implementation is not justified, such as those with an extremely low accuracy that directly affect the wellbeing of a population group. In these cases, the option of not implementing shall weigh in the discussion about its convenience.

When implementation is essential and implies a bearable risk, it is necessary to take mitigation and control measures to reduce the errors and their effects. Here, there are different options: from constant reviews of the system's results to human intervention in the decision on a specific social subgroup or all participants to algorithmic transparency, among others.

It is worth noting that there will also be cases in which audits fail to meet their goals because they were conducted at less than optimal times, namely:[12]

- **Premature audits:** In these cases, audits have been conducted at a very early stage, before significant aspects of the system have been fully implemented or before it is possible to correctly assess the possible damages.
- **Late audits:** Being an ex-post exercise, audits can remedy future situations, but not address early damages. Often situations that caused damage for a considerable period are detected, without these being evaluated and/or mitigated timely.
- **Sporadic audits:** Audits are an ongoing developing mechanism that evolves and matures according to advances in technology and society. Therefore, potential impacts anticipated in an audit performed during the first months of implementation of the system may differ from those in later stages. It is, therefore, highly recommended to carry out these procedures periodically to keep the audit mechanism and the information provided by the entity involved up-to-date.

2.4 Who should perform an algorithmic audit?

Audits can be external and internal to assess the effectiveness and potential consequences of the system. Depending on who performs it, there can be three types of audits:[13]

- **Third-party audits:** External agents evaluate the behavior of a system based exclusively on its results.
- **Second-party audits:** A supplier, client, or contractor of the audited institution is granted access to the system server (backend) and evaluates its behavior considering the technical aspects and the results.
- **Internal audits:** A member or team of the entity evaluates the entity's concerns. Typically, such alarms originate from the usual challenges of the responsible development of AI systems, such as transparency and equity. These audits seek to reach goals related to the system itself, considering its success criteria.

It is worth noting that in the three categories above, and regardless of whether the auditor is internal or external, an imperative and common requirement is that whoever is in charge of conducting the audit should not be involved in the system development.

It is possible to perform more than one audit in more than one of the types indicated; in such cases, it is important not to duplicate errors when not justified and ensure that audits complement each other.

The audit team competencies should include (i) technical expertise, (ii) knowledge about the specific area of ADS implementation, and (iii) robust knowledge about the ethical principles that must be incorporated in standalone systems.

A technically expert team means that they are proficient in specific programming languages and methodologies used in ADS, and can validate the data selection and work. For example, if someone calling themselves an auditor is not a specialist in the AI technology used in the entity's ADS, little can they decide on the auditable compliance of that system.

Knowledge about the specific ADS area of implementation is critical for the audit team to evaluate and make recommendations about improvements to the system. For example, in a system that evaluates the dangerousness of individuals undergoing criminal proceedings (see COMPAS case in Section 5) and considers the skin color of the charged individual and/or the neighborhood where they come from – which is contrary to human rights per se -, auditors must

12 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

13 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-It interest/

be able to notice the bias if they want to fulfill the purpose of improving the system.

The audit team must have a **robust knowledge of the ethical principles that must be incorporated in the autonomous system** to be able to evaluate them during the review. Today, there are different ethical frames allowing this, as shown later. It is also necessary to adopt data protection principles and tools to prevent or reduce cases of discrimination. The audit team should determine whether the ethical rules and principles that govern the ADS operation protect the dignity of individuals rather than focusing solely on efficiency or system failures.

From the audit design perspective, communication channels must be established between the team responsible for the audit and the audited institution, paying special attention to the positions and functions related to the development of the system described in paragraph 2.10.

2.5 What information is shared in an audit process, and how?

A system audit requires a record of detailed documentation about the training processes, the performance and validation of tests, and their implementation. The more detailed the documentation, the easier the auditors' job will be. However, the cost and efforts invested in documenting the lifecycle of the standalone system will depend on the criticality level of the process.

Detailed documentation allows auditors to review the development history of the algorithm, including its original purposes, the participating team, the tests performed, and the modifications it has undergone. Accordingly, it is possible to compare the system stages, which is extremely useful in determining exactly when an abnormality might occur.

2.6 What should the documentation include?

An algorithmic audit implies several evaluation processes. These range from the entity's governance model whose system will be audited (organization chart and functions of the team involved in its development, strategic plans for its use and implementation, stakeholders and affected parties, among other elements), the databases used (data collection methods, and quality, pertinence, management, and handling of these, among other elements), to the computational method (algorithms used, system sensitivity and specificity). Therefore, the documentation delivered by the audited company shall allow the understanding of the system's governance model and build an appropriate profile of the data[14] and the algorithmic model itself[15].

To build the profile of the data used in the model, the information of its source, collection, governance, and structure is necessary, together with an evaluation of its quality. To build the model's profile, it will be necessary to have information about its conceptualization and design; the source and handling of the data; its development, use, and monitoring, and the relevant accountability. The document "Responsible use of AI for public policies: Data science toolkit"[16] offers a detailed approach of the items contained in each of them.

2.7 In what conditions should data be delivered to the auditor?

Depending on the system criticality level, it will be important to define the conditions for the transfer of documentation. For this, the permitted and prohibited uses of the algorithms and the data must be explicitly defined. A Data Sharing Agreement will be very useful for this purpose, establishing the party's roles and responsibilities; clarifying the purpose of the data transfer, detailing what happens with these in each stage, and establishing the use, security,

14 IDB. Responsible AI: Technical Manual: Life Cycle of Artificial Intelligence (IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial), p. 55. Available in Spanish at https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

15 IDB. Responsible AI: Technical Manual: Life Cycle of Artificial Intelligence (IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial), p. 57. Available in Spanish at https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

16 Ibid.

and privacy standards.[17] Accordingly, both the auditor and the auditee will have a document listing each party's data responsibilities, which is particularly relevant, for example, in cases where confidentiality is critical, such as those involving personal data, national or public security information, or commercially sensitive information.

Training data may be delivered to the auditors for them to reproduce the process and evaluate whether there is a better way to work with such information. To perform this procedure, complying with the regulations on personal data protection will be essential, if applicable. For example, it would be convenient to anonymize the database to fully protect such information. Also, the transfer agreement must specify that these data will be used for auditing as an exclusive purpose and that they will not be used for other purposes.

2.8 Who should have access to the results of the audit?

The general rule of public administration is the **principle of transparency**, according to which the acts, resolutions, procedures, and documents of the state administration must be public. This principle allows the State to be accountable to civil society, represented directly by social or community organizations, universities, and think tanks.

In the case of the results of an algorithmic audit, it is important to determine who are the third parties that will have access to the information and the evaluation prepared by the auditors. The audit report will contain an analysis of the algorithms' efficacy but may also show its operation, the types of data used, and the possible vulnerabilities, so it is therefore critical to analyze and determine the extent of disclosure of the reports.

To this effect, the information of the automated process that is sensitive and that which can be freely known by third parties must be first identified. The next step is to identify which of the elements subject to analysis by the auditors may be a risk for the operational continuity of the system and the protection of beneficiaries. Lastly, it should not be ignored that feedback from the public allows improving procedures in directions not necessarily anticipated, and democratically reviewing the processes that have an impact on people's lives. Feedback from citizens and civil society can be done directly, with at least an e-mail address to process complaints and suggestions. Transparency is not only fulfilled by disclosing information, but also through participation mechanisms in which individuals can express their concerns directly to the authority.

The sensitivity of the information will depend on whether or not its knowledge may cause harm to the beneficiaries, adversely affect the system's operational continuity, or affect the effectiveness of the service. On the contrary, biased information about certain groups or evidencing discrimination should not be considered sensitive and/or confidential. It should be made public so that beneficiaries can defend themselves and protect their interests if errors have occurred.

Depending on the need to maintain confidentiality, audits can be classified according to their degree of transparency and disclosure. A case of maximum confidentiality (less disclosure) would be that in which the audit results are only known by the agency that is implementing the system. A medium confidentiality case would be that in which the same information can be shared with higher hierarchical agencies or evaluators. A lower confidentiality case would occur when the information used can be shared with peer public entities that can benefit from it to improve their processes.

The minimum confidentiality level is when information is shared with third parties outside the public administration, such as international institutions, universities, or think tanks. Here, the data integrity can be assured by verifying that the shared information is governed by Non-Disclosure Agreements, guaranteeing thus that it is not disclosed or used for purposes other than the audit.

17 UK Information Commissioner's Office. Data sharing code of practice. Available at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/ico-codes-of-practice/data-sharing-a-code-of-practice-1-0.pdf

Finally, information will be widely disclosed when its knowledge does not imply danger to the operational continuity of the system, to individuals, and/or to the efficiency of the service. It is also possible that the audit becomes public after neutralizing the identified risks and remediating the vulnerabilities.

2.9 Considerations to perform an algorithmic audit

The underlying assumptions to conducting an algorithmic audit include the following:

(i) Auditor, whether internal or external, is independent and external to the development of the system implementation.

(ii) Whoever develops and implements the system must be able to supply appropriate information to the auditor.

(iii) Auditor must correctly understand the system based on the information provided, the relevant documentation, and the effects it may perceive regarding the system's impacts.

(iv) The system behavior during its use and monitoring is consistent with its behavior when audited[18]. It is imperative to consider this, as its performance may vary depending on the context or the data feed. The latter is what justifies the need to conduct these audits periodically to recognize possible changes in the scenario, include of new functions, or remove others.

(iv) Whenever possible, audit should be conducted in binary terms, meaning that evaluations should be in a format that does not allow for nuances (e.g., compliant/non-compliant). The reason for this is that an evaluation ranking may lead to grey areas that undermine the clarity and confidence required in the audit.

2.10 Damage determination

During the audit, the damage generated by a system due to a defective, imperfect, or suboptimal operation will be revealed – and can be measured.

Damages are losses suffered by beneficiaries and third parties external to the system. It will always be necessary to correctly define each affected group, clearly describing the features of each one. It will help recognize present patterns or qualities that can be subject to greater or lesser scrutiny than what is considered optimal in a context where the system is used.

Examples of affected groups

It is important to keep in mind that the groups affected by ADS are not only those who use them or are directly involved in the actions or recommendations arising from such systems. Often there are unconsidered third parties outside the systems affected by their use.

Take, for example, a system that determines the frequency of public transport and whose decisions affect clearly and directly its users by indicating whether or not there should be a higher frequency in a given period of the day to optimize the use of public resources and user satisfaction. However, other groups are also affected, like private vehicle users, pedestrians, and cyclists, as their travel times would also be affected by this frequency.

18 Ada Lovelace Institute. Algorithmic Accountability for the Public Sector. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

As indicated in the System Criticality box, damages may seriously affect two types of key factors: (i) those that directly influence the allocation or restriction of rights, aids, or subsidies, and (ii) those that are part of a chain of processes, and which, in the event of failure or error, can affect any of the elements that impact on the provision of government services. Cyber-attacks have been particularly aimed at the latter, affecting some State services.

The European Union has defined four risk levels for the IA models:

(i) **Unacceptable risk:** Those applications considered harmful to the health and integrity of individuals and that contravene fundamental rights. These are prohibited.

(ii) **High risk:** Applications that harm the safety of people or those that are safety components of larger systems. These must be evaluated by third parties and compliant with the sector regulations before coming into operation.

(iii) **Limited risk:** Applications with a low-risk level that must comply with transparency and information requirements for those citizens who are subject to automated processing.

(iv) **Minimal risk:** Any system whose application does not imply any risk. Developers may voluntarily adhere to codes of conduct.

Unfortunately, when dealing with ADS systems, it is not always possible to obtain an explanation about the error's reason and cause and its consequent damage. It is known as the "black box" problem. Technical audits are, therefore, limited to evaluating whether or not the necessary precautions were taken when the system was developed.

Black box

The "black box" metaphor applies to systems where the internal mechanisms are unknown; either because they are impossible to understand or because doing so is very expensive and is unreasonable (E.g., trying to understand a neural network). In such cases, it is impossible for a human being to discern how some inputs (E.g., data) lead the system to produce a result (E.g., a particular action or recommendation).a

Although better performance may justify using these models, it opposes the search for transparency in ADS implementation.

[a] Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Available at: https://www.auditingalgorithms.net/

2.11 Profiles and functions of audit collaborators

The implementing entity's contact officials and their functions shall be available; they are responsible for providing the requested information if clarification or further information of the system background subject to audit is required. The following is a description of the positions and functions that may exist for these systems in an average entity:[19]

» **Chief Information Officer (CIO):** Person in charge of an entity's IT systems. Decides and directs the IT developments to achieve the institution's goals.

» **Chief Privacy Officer (CPO):** Person in charge of making the institution's privacy decisions and protecting the interests of the beneficiaries in this area.

19 While smaller entities may only have one person in charge, larger entities may have entire teams dedicated to fulfilling the tasks of a particular function.

- **Chief Information Security Officer (CISO):** Person in charge of the security of the information produced and possessed by the entity.
- **Legal Director:** Person in charge of the institution's legal affairs and compliance.
- **Software Developer:** Person in charge of programming the system and converting the institutional requirements into software compliant with the desired technical purposes.
- **Data Analyst:** Person in charge of analyzing, organizing, and debugging data to serve as an input for decision-making within the institution.
- **Data Engineer:** Person in charge of building and maintaining the databases and preparing them for later use by the Data Analyst.
- **Project Director:** Person in charge of the project execution, maintaining its cohesion, and distributing tasks within the team.
- **Product Owner:** Person in charge of the system development practical tasks related to strategy, execution, and launch.
- **Expert:** Person with profound knowledge who can contextualize users' needs.

3. Stages of an Algorithmic Audit

While there is currently not a unique model for algorithmic audits, this guide uses the Raji & Smart, et al. (2020), which comprises six stages: (i) definition of the audit scope, (ii) stakeholders mapping, (iii) documentation collection, (iv) testing, (v) results analysis, and (vi) post-audit[20].

The following are the tasks that should be completed in each of the six stages mentioned above:

» Definition of the audit scope

- Product Requirements Document
- Review of principles considered in the system design
- Analysis of similar use cases
- Social impact assessment

» Stakeholders mapping

- Definition of questions and team interviews
- Responses transcripts and systematization

» Documentation collection

- Preparation of audit checklists
- Preparation of data profiles
- Preparation of the model profile

» Testing

- Review documentation
- Faults simulation and search for vulnerabilities
- Preparation of the system use risk matrix

» Results analysis

- Risk matrix update and formalization
- Preparation of an action or risk mitigation plan
- Compilation of detail and evolution of system development
- Audit report

20 Raji et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Available at https://arxiv.org/pdf/2001.00973.pdf

Note that these steps will not always be sequential, and it is possible that from the early stages of the audit, the system may be found to be unfeasible, making it unnecessary to move forward to later stages.

Guiding questions to conduct an audit correctly are included in the annex.

To date, the various existing AI regulations in Latin America and the Caribbean do not refer directly to the algorithmic responsibility, unlike counties with more mature jurisdictions in the development of such topics as Canada,[21] Sweden,[22] or the United Kingdom[23].

Given the growing number of legal initiatives impacting the development of standalone systems, as well as policies about the use of AI in the region, the response will also be determined by compliance with the legislation and the policies that seek to ensure the appropriate use of the system and the specific industry regulations. It is necessary to take into account that there are related areas that will have an impact on the review of the compliance standards, for example, in data protection, cyber security, anti-discrimination laws, or even sectorial regulations.

Regarding the public sector, currently, there is not a standardized practice for conducting algorithmic audits. However, there are some initiatives targeting the consolidation of experiences in various cases, areas, and jurisdictions. An example is the document "A White Paper for Public Auditors",[24] prepared by the audit authorities of Finland, Germany, the Netherlands, Norway, and the UK based on their experience.

Following the audit, the entity must determine whether it is possible to continue using the system or if it should be partially or totally modified, according to the responses obtained (classified in the annex according to their relevance: extreme urgency, extreme importance, or recommended revision). If required, the action or risk mitigation plan should be implemented, together with a subsequent ongoing follow-up of its implementation.

21 As an example, in Canada the 2019 directive on automated decision-making aims to reduce the risks of these systems, and achieve more efficient, accurate, consistent, and interpretable administrative decisions under Canadian law. Accordingly, audits implementation is expanded, access to information is facilitated, and the data quality standard is raised. Directive on Automated Decision-Making. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

22 Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up. Available at: https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

23 Guidance on the AI auditing framework, Draft guidance for consultation. Available at: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

24 Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Available at: https://www.auditingalgorithms.net/

4. ADS guiding principles

To encourage the implementation and ethical use of IA-based systems, varying jurisdictions and organizations have adopted guiding principles, both partially and for the entire system. For example, there are the principles of Article 5 of the European Union General Data Protection Regulation (GDPR)[25] that specifically regulate the processing of personal data. For the system as a whole, there are the principles set out in the Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI[26] of the Berkman Klein Center for Internet and Society at Harvard University.

This guide will consider OECD's list of ethical principles included in the Recommendation of the Council on Artificial Intelligence.[27] It is the first set of intergovernmental policies on AI composed of the principles translated in the document Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean,[28] which are summarized below:

Inclusive Growth, Sustainable Development, and Well-Being. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet. The appropriate use of AI can promote the augmentation of human capabilities and enhance creativity, advance the inclusion of underrepresented populations, reduce economic and social inequalities, and protect natural environments, thus invigorating inclusive growth, sustainable development, and well-being.

Human-Centered Values and Equity: AI actors should respect the rule of law, human rights,

25 https://gdpr-info.eu/art-5-gdpr/

26 Fjeld & Nagy. Principled Artificial Intelligence. Available at: https://cyber.harvard.edu/publication/2020/principled-ai

27 OECD. Recommendation of the Council on Artificial Intelligence. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

28 IDB, fAIr LAC. Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean. Available at: https://publications.iadb.org/publications/english/document/fAIr-LAC-Responsible-and-Widespread-Adoption-of-Artificial-Intelligence-in-Latin-America-and-the-Caribbean.pdf

and democratic values, throughout the AI system lifecycle. These include freedom, dignity, and autonomy; privacy and data protection; non-discrimination and equality; and diversity, fairness, social justice, and internationally recognized labor rights. To this end, AI actors should implement mechanisms and safeguards, such as the capacity for human determination. These must be adjusted to the context and consistent with the state of art.

Transparency and Explicability. ADS must allow the stakeholders of the ecosystem to understand their functioning and possible outcomes. Therefore, implemented systems should be governed by the principles of transparency and responsible and fair disclosure of information.

Relevant information should be provided to those who use the system and to passive subjects of the analysis. Information should be tailored to the contexts of the recipient of the information, in such a way that the recipient can fully and correctly understand it.

The objectives are: (i) to foster a general understanding of AI systems; (ii) to make stakeholders aware of their interactions with AI systems; (iii) to enable beneficiaries and passive subjects to understand the potential outcomes and risks of using ADS; and, (iv) to enable those adversely affected by an AI system to challenge its outcome based on clear and easy-to-understand information on the factors and the logic that served as the basis for the prediction, recommendation or decision.

It is critical that decision-makers also understand the operation and potential risks of the use of ADS, to incorporate their analysis where the machine can fail or present risks. As described below in Section 5 of this document on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)[29] case, the objective was to identify the risk of re-offending by individuals who had been prosecuted.

The use of COMPAS made headlines as it showed a favorable bias towards white people and an adverse bias towards darker-skinned individuals. Something similar happened in the case of felonies committed by men and women, the latter being the most punished.

Given that none of the passive subjects of the COMPAS system knew its operation because it lacked transparency, the judges who relied on it did not question its recommendations. This only happened later, following a press report, which resulted in the system being no longer used.

Had the system been transparent, it would have been evident that it was considering elements that are not specific to a sanction, such as ethnic origin, family composition, and/or education level of the defendant. Likewise, the latter would have had the opportunity to defend themselves from the penalties imposed based on the ADS, as they were clearly contrary to due process.

The transparency of the systems not only allows passive subjects or recipients of their actions to exercise their rights, but it also helps decision-makers to weigh and analyze the validity of the recommendation, so that they fully understand it, and determine whether it constitutes an element that respects the dignity of individuals, human rights, and the rule of law.

Robustness, security, and safety. These are three essential elements of every AI system for the following reasons:

» AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use, or misuse, other adverse conditions, they function appropriately and do not pose an unreasonable safety risk.

» To this end, AI actors should ensure the ongoing traceability of datasets, processes, and decisions made during the AI system lifecycle, allowing to analyze the outcomes and responses to inquiry correctly, and consistently.

» Based on their roles, the context, and their ability to act, AI stakeholders should apply a systematic risk management approach

29 Brennan, T. y Dieterich, W. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Available at: https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.

to each phase of the AI system lifecycle continuously to address risks related to AI systems, including privacy, digital security, safety, and bias.

Accountability. AI actors should be accountable for the proper functioning of AI systems and the respect of the above principles, based on their roles, the context, and consistent with the state of art.

5. Use case: ADS in predictive surveillance

To highlight the relevance of making an algorithmic audit, its use in the context of predictive policing is examined below.

In the United States, the use of ADS has been implemented in the risk analysis for the criminal defendants community in multiple states. The system used, developed by Northpointe is called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The system provides a score to the relevant court based on the answers to a questionnaire of 137 questions along these lines: Has your father/mother ever been in prison? How often did you get into fights at school? These questions were either responded to by the defendants or obtained from their criminal records.

In 2013, Eric Loomis was arrested for driving a vehicle carrying persons that has recently participated in a shooting. According to the recommendation made by the COMPAS system, indicating that he was a highly dangerous individual to the community, he was sentenced to six years of imprisonment and five years of extended supervision.[30]

In this case, several issues come to light, regarding the suitability of the use of the system in the context mentioned above, its accuracy, and the bias that the data used may have, among others. Therefore, before using ADS in such complex contexts as the administration of justice, it is necessary to always respond to the following questions, which will use the COMPAS case as an example.

A clear purpose for the use of the system has been defined?

How is it ensured that the system is not used for purposes other than those for which it was developed?

Regarding these questions, in the COMPAS case, Tim Brennan, founder of Northpointe, said that its focus in designing this system was to reduce crime, and not to determine penalties. As described above for this case, the system deteriorated and ended up being used as a basis for determining the guilt of the defendant, which is far from the original purpose for its development.

30 Wisconsin Supreme Court. *State v. Loomis.* Available at: https://harvardlawreview.org/2017/03/state-v-loomis/

Has the system been tested in different demographic groups to mitigate the existing biases?

Have measures to mitigate historical biases in the databases used been taken?

The fact that the COMPAS questionnaire was made up of questions about the defendant's childhood, ancestors, or neighborhood should have raised alarms about potentially biased data. In this case, appropriate measures to mitigate the existing historical biases in the data used were not taken, which resulted in an erroneous risk score allocation to individuals with disparate background checks.

Are the definition of the architecture and the techniques used consistent with the needs for transparency and explainability of the decisions required by the sector in which the system is embedded?

In certain sectors, the explainability of decisions is critical for an appropriate acceptance of the AI systems by society. In terms of justice administration, explainability is essential. In the case mentioned above, the interested parties did not know how the score was calculated, because Northpointe maintained such information as a trade secret.

Considering the featured instability of several automated learning models, has the model been validated on multiple occasions and scenarios to ensure that it responds correctly in varied contexts?

How have the optimal sensitivity and specificity points been defined in the ROC curve?[31] Are these adequate for the industry in which the system will be implemented?

A subsequent system evaluation that analyzed 16,000 cases revealed that its accuracy was close to 71%.[32] Since the implementation context – justice administration – is quite sensitive, clearly the accuracy is far from what could be considered ideal. Consequently, longer testing periods and higher validations parameters were required.

31 The ROC curve (Receiver operating characteristic) is a statistical tool to assess the accuracy of a model's predictions. E.g., If a model that classifies people according to their risk of committing an offense is going to be implemented, the ROC curve may accurately evaluate such a model.

32 Angwin et al. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

6. Final remarks

AI has a critical role in our day-to-day and in our coexistence as a society, to the point that it becomes increasingly difficult to even think on an instance in which today, we do not interact with intelligent systems. Mobile devices, household appliances, means of transportation, among many others, make our activities easier, more comfortable, and safer.

As the AI technology, the algorithmic audit arena advances at an accelerated pace, increasing the importance of its use. It is constantly evolving with a scope in permanent flux. Therefore, its contents shall be updated regularly according to the development of new IT tools and the corresponding regulations.

Given the massification of ADS in society, and in particular, in areas that require extraordinary precautions for its use, it is necessary to make ongoing revisions to guide their correct implementation. Cases of systems that due to design or development flaws have had significant adverse impacts on our daily lives and society have already happened, ranging from increased public transportation fares to unfair convictions.

One of the challenges for the public sector is the transition from conducting algorithmic audits as a voluntary mechanism to including these either as part of a structured policy on the matter or as part of a widespread regulation on algorithmic liability.

This guide is not only intended to be a practical instrument to monitor critical areas and mitigate risks or hazards that may not be evident to the naked eye. It is also expected to serve as a tool to help in raising awareness about the implications and consequences of implementing ADS. We hope that all the teams, both from public entities and ADS developers are aware of the relevance of their work. If we want to guarantee a more just and safe future, we must understand its relevance and make it fully understood.

References

Ada Lovelace Institute (2021). Algorithmic Accountability for the Public Sector. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

---------- (2020). Examining the Black Box: Tools for assessing algorithmic systems. Available at https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Angwin, J., Larsson, J., Mattu, S. y Krichner L. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

IDB (2021) Responsible use of AI for public policy: A project formulation guide. Available at: https://publications.iadb.org/publications/english/document/Responsible-use-of-AI-for-public-policy-Data-science-toolkit.pdf

IDB Lab (2021) Ethical Assessment of AI for Actors within the Entrepreneurial Ecosystem: Application Guide. Available at: https://publications.iadb.org/publications/english/document/Ethical-Assessment-of-AI-for-Actors-within-the-Entrepreneurial-Ecosystem-Application-Guide.pdf

IDB, fAIr LAC (2020). Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean. Available at: https://publications.iadb.org/publications/english/document/fAIr-LAC-Responsible-and-Widespread-Adoption-of-Artificial-Intelligence-in-Latin-America-and-the-Caribbean.pdf

Brennan, T. y Dieterich, W. (2017). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Available at: https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.

Fjeld, J. y Nagy, A. (2020). Principled Artificial Intelligence. Available at: https://cyber.harvard.edu/publication/2020/principled-ai

Government of Canada. (2021). Directive on Automated Decision-Making. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

IA Now Institute (2028). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, available at: https://ainowinstitute.org/aiareport2018.pdf

Intersoft Consulting. S.f. Data Protection Regulation. Available at: https://gdpr-info.eu/art-5-gdpr/

INTOSAI (2019). Performance Audit Principles. Available at https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

ISO 19011-2018 (2018). Available at https://www.iso.org/standard/70017.html

Moss, E., Watkins, E. A., Singh, R., Elish, M. C. y Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Available at https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

OECD (2019). Recommendation of the Council on Artificial Intelligence. Available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD/IDB (2020) Responsible AI: Technical Manual: Life Cycle of Artificial Intelligence (IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial). Available in Spanish at: https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

Raji, I. D., et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Available at https://arxiv.org/pdf/2001.00973.pdf

Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Available at: https://www.auditingalgorithms.net

Swedish National Audit Office (2020). Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up. Available at https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

UK Information Commissioner's Office (2020). Guidance on the AI auditing framework, Draft guidance for consultation. Available at https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

US Congress (2019). Algorithmic Accountability Act of 2019.


IDB
Improving lives